Глава 4. Как велика ошибка, когда мы прогнозируем колебательный спектр молекулярной модели?

Хороший вопрос. К сожалению, никто не может дать столь же хороший ответ. Настолько хороший, чтобы он удовлетворил любого прикладника, пользующегося теорией колебаний молекул при решении практических задач. Всем приходится довольствоваться весьма уклончивыми ответами, либо делать вид, что такой вопрос задавать вообще неприлично, раз эта теория оперирует столь сложными моделями. Во всяком случае, я знаю, что ни в одном руководстве по моделированию колебательных спектров молекул этот вопрос не ставится в той категорической форме, как здесь его ставлю я. И отвечаю на него в меру моих сил и возможностей.

Что такое точные науки? С таким вопросом ко мне, когда я преподавал физику в школе, неожиданно подошла ученица 11 класса. Девочка была красивая, мне не захотелось быть скучным (стандартный ответ – Это математика, физика, астрономия и многие инженерные науки, в частности, артиллерия, баллистика). Я тут же сочинил нестандартный ответ. Это такая наука, которая сама может узнать и сказать нам, насколько она неточна в своих прогнозах. Девочка тут же удалилась. Развивать и раскрывать оригинальную мысль не пришлось. Осталось чувство неловкости. И застряло желание сделать мою науку, для которой я как программист готовлю рабочий инструментарий, точной наукой.

4.1. Что мешает просто и ясно ответить на этот вопрос, в чем тут проблемы?

Помех и проблем в случае прогнозирования колебательного спектра молекулы две:

1. сложность (многомерность) моделей;
2. неопределенность статистических законов распределения многочисленных параметров, используемых в моделях.

Многомерность модели мешает дать точный рецепт вычисления погрешности

Пусть имеется модель некого явления и содержательная теория, позволяющая вычислить характеристики этого явления как функцию параметров модели и некоторых глобальных параметров. С математической точки зрения это означает, что определена функция

$$y = f(x), \qquad (1)$$

где x есть вектор параметров модели размерности n, а y есть вектор выходных характеристик размерности m.

Функция (1) не обязательно должна быть представлена в виде формулы. Это может быть алгоритм, построенный на основе данной теории, и реализованный в виде вычислительной процедуры.

Задача заключается в том, чтобы наряду с вектором y вычислить вектор погрешностей прогноза Δy. Необходимо также определить статистические свойства компонент случайного вектора y на основе ранее исследованных статистических свойств компонент случайного вектора x. Это даст возможность построить доверительные интервалы для всех прогнозируемых характеристик данного явления в форме

$$y \pm t \;\; Q \;\; \Delta y, \tag{2}$$

где Q есть заданная доверительная вероятность, а t – множитель, аналогичный коэффициенту Стьюдента в одномерной статистике. Может оказаться, что Q и t будут векторами размерности m, но мы пока не будем касаться этого усложнения задачи.

Заметим, что математическая статистика не дает общего рецепта построения выражений (2). В теории строго доказаны соответствующие теоремы только для случаев

$$y = \sum_{i=1}^{n} c_i x_i; \quad y = \sum_{i=1}^{n} x_i^2; \;\; m = 1.$$

При этом статистические свойства выражений (2) прояснены только, если каждый компонент вектора x распределен нормально. Поэтому на практике в каждой области знания специалисты находят эвристические рецепты построения выражений (2). В частности, в физических исследованиях предлагается совершать следующую логическую подмену в случае $m = 1$ (случаи с $m > 1$ вообще не рассматриваются).

Найдем частные дифференциалы от функции y по всем измеряемым в эксперименте параметрам x_i модели, и будем считать, что свойства квадратов этих частных дифференциалов аналогичны свойствам квадратов случайных погрешностей. (Логическая подмена, строго не обоснованная, состоит в том, что в известные формулы математической статистики подставляют эти дифференциалы вместо дисперсий независимых величин. Пользуясь этим эвристическим приемом, получают рабочие формулы для оценок погрешностей). По аналогии с точным выражением для дисперсии суммы двух статистически независимых величин x_1 и x_2

$$s^2 = s_1^2 + s_2^2, \tag{3}$$

где s есть стандартное отклонение случайной величины, получим (при $m = 1$)

$$\Delta y = \sqrt{\sum_{i=1}^{n} \left(\frac{dy}{dx_i} \right)^2 \Delta x_i^2}. \tag{4}$$

Возможность указанной логической подмены в теории статистики не доказана, статистические свойства формулы (4) в общем случае не прояснены, поэтому за простоту выражения (4) приходится расплачиваться. На практике в выражение (4) подставляют максимальные оценки погрешностей параметров, что дает преувеличенную ширину доверительного интервала для прогноза величины y при $Q = 100$ %. Это сразу же исключает возможность надежно регистрировать и интерпретировать малые изменения y при малых вариациях условий опыта. В физических экспериментах, когда выражения для

y достаточно просты, с этим можно мириться. Однако в случае многомерной модели, когда *n* велико, выражение (4) может дать недопустимо большую оценку ширины доверительного интервала. В то же время, оценка надежности $Q = 100$ % никому практически не нужна. Исследователь должен иметь возможность гибкого управления соотношением между шириной доверительного интервала и приписываемой ему надежностью. В полном соответствии с принципом дополнительности Бора.

Реплика в сторону. В теории колебаний молекул разработаны способы вычисления производных от частот по силовым параметрам модели. Этим способам соответствуют очень простые вычислительные программы. Однако из вида формулы (4) сразу понятно, что использовать эти производные для оценки ошибки прогнозирования частоты колебаний совершенно нерационально. Частота нормального колебания зависит практически от всех силовых постоянных модели, и сумма квадратов частных дифференциалов под корнем в формулы (4) получается недопустимо длинной даже для небольших по размеру молекул. Производные от частот по силовым параметрам имеет смысл вычислять, когда надо прояснить зависимость частоты от какой-то одной силовой постоянной. Мы будем так поступать при постановке и решении обратных спектральных задач.

Вычислительная процедура оценки точности прогноза для сложной многомерной модели

Предлагается следующее общее решение данной проблемы, основанное на интенсивном использовании вычислительной техники. Способ использования вычислительной техники похож на известную процедуру бутстрепа.

Соберем и оформим в виде законов распределения всю возможную информацию о статистических свойствах отдельных компонент вектора *x* параметров модели. Эта информация может быть известна из документации к используемым в эксперименте приборам. Это может быть информация, полученная в результате решения обратных задач после проведения специальных экспериментов. Это может быть обоснованное предположение о законах распределения соответствующих случайных величин. Важно только, чтобы были определены вычислительные процедуры, позволяющие генерировать случайные значения компонент вектора *x* в соответствии с принятыми законами их распределения. Затем выполним множество прогнозов для конкретных вариантов вектора *x*, используя имеющуюся процедуру для вычисления выражения (1). При этом будем каждый раз генерировать случайную реализацию вектора *x*, пользуясь приготовленными ранее процедурами получения случайных значений всех компонент этого вектора. Мы получим множество значений для каждой из *m* компонент вектора *y*. Считая все компоненты вектора *y* статистически независимыми, построим для каждой компоненты этого вектора свою гистограмму. Это даст возможность для каждой компоненты задать требуемую надежность Q_j и по гистограмме непосредственно оценить границы соответствующего доверительного интервала для значения характеристики y_j исследуемого явления. При этом совершенно не требуется выводить и анализировать законы распределения для отдельных компонент вектора *y*, поскольку все их статистические свойства будут непосредственно видны из полученных гистограмм. Этого вполне достаточно для разработки конкретной методики использования прогноза в любой научной деятельности.

Если может проявиться статистическая зависимость между компонентами вектора *y*, то можно выявить эту зависимость, построив выборочное представление ковариационной матрицы, используя данные о множестве вычисленных реализаций вектора *y*. Это позволит внести нужные коррективы в полученные представления доверительных интервалов для компонент прогнозируемого по закону (1) вектора *y*.

Предложенная вычислительная процедура аналогична бутстрепу и является его обобщением на случай сложных многомерных моделей. Существенным отличием здесь является использование информации о законах распределения параметров, а не экспериментальных данных непосредственно, как это делается в оригинальной процедуре бутстрепа. Можно утверждать, что применение бутстрепа при решении практических задач тем более эффективно, чем более сложной является функция от случайных величин. Такое утверждение было высказано в [1].

Единственное слабое место данного предложения связано с необходимостью выполнения большого объема вычислений. Если вычисление выражения (1) требует заметного машинного времени для выполнения единичного прогноза, то оценка величин доверительных интервалов для всех компонент прогнозируемого вектора *y* потребует в 10 – 100 – 1000 раз больше времени, в зависимости от представлений исследователя о необходимой скрупулезности предпринимаемой работы. Однако эта особенность предложения не является серьезной угрозой. Приведем конкретные оценки требуемого машинного времени для самых сложных моделей, с которыми приходилось иметь дело в конкретных исследованиях с применением теории колебаний молекул. Расчет полного вида спектральной кривой ИК поглощения для крупной модели органической молекулы в несколько десятков атомов требует не более 20 секунд на настольном персональном компьютере. Достаточно полные представления о виде гистограмм распределения всех частот и интенсивностей в ИК спектре такой молекулы потребуют не более 20 минут машинного времени. В процессе разработки методики использования расчетных колебательных спектров для решения конкретной научной задачи такое время вполне можно потратить.

Если же потребуется более скрупулезное представление о поведении частот и интенсивностей при варьировании параметров модели, то на суперкомпьютере МВС-1000 можно одновременно запустить до 1000 вариантов одного и того же расчета с разными наборами параметров модели. И все нужные результаты будут получены за те же 20 -30 минут. Выход же не суперкомпьютер указанного класса сегодня доступен любой лаборатории.

Демонстрационный пример – простая двумерная модель

Рассмотрим простой случай, когда $n = 2$, $m = 1$.

Пусть по результатам измерений требуется оценить площадь прямоугольника *S*. В соответствии с формулой (1), для данной простой модели $y = S$; вектор *x* содержит две компоненты: $x_1 = a$ – длина прямоугольника, $x_2 = b$ – ширина прямоугольника. Рабочая формула для прогноза площади прямоугольника

$$S = ab. \quad (5)$$

Пусть нам известны законы распределения независимых случайных величин *a* и *b*. Закон распределения случайной величины *S* определяется композицией этих законов. Известно, что аналитическое выражение находится легко только для композиции двух нормальных распределений в случае суммы случайных величин. Для других законов распределения, для других функций получить такое аналитическое выражение уже трудно. Таким образом, строго доказанной формулы для закона распределения функции (5) мы не имеем. Общепринятый рецепт построения доверительного интервала для величины *S*, в соответствии с (4), дает формулу

$$\Delta S = \sqrt{a^2 \Delta b^2 + b^2 \Delta a^2}. \quad (6)$$

Свойства случайной величины ΔS с трудом поддаются интерпретации. Если под Δa, Δb понимать максимальные погрешности измерений, то ΔS есть максимальная погрешность прогноза ширины прямоугольника с надежностью $Q = 100$ %. Если же нас не устраивает такая надежность и соответствующее завышение неопределенности результата, то под Δa, Δb можно понимать стандартные отклонения измеряемых величин. Но тогда мы ничего не знаем о надежности доверительного интервала для прогноза.

Используем предложенную выше методику. Пусть величины *a* и *b* распределены нормально с дисперсиями σ_a^2, σ_b^2. По результатам измерений мы можем оценить эти неизвестные параметры через соответствующие выборочные дисперсии s_a^2, s_b^2 и соответствующие среднеквадратичные отклонения s_a, s_b. Пусть мы имеем для средних значений

$a = 5$, $b = 2$

и для среднеквадратичных отклонений

$s_a = 0.2$, $s_b = 0.1$.

Используем программу Rectangle, написанную на языке MatLab для реализации описанного алгоритма. Программа снабжена генератором нормально распределенных чисел с заданным среднеквадратичным отклонением. Программа создает выборку бутстрепа для случайных значений величины *S*. При подсчете каждой конкретной величины S_i программа использует независимо сгенерированные величины a_i, b_i. Индекс *i* пробегает в программе значения от 1 до $n_b = 10000$. Программа строит гистограммы величин *a* (*b* имеет аналогичную гистограмму) и *S*, а также подсчитывает значения случайных реализаций величин s_a, s_S. Гистограммы показаны на рисунке 1.

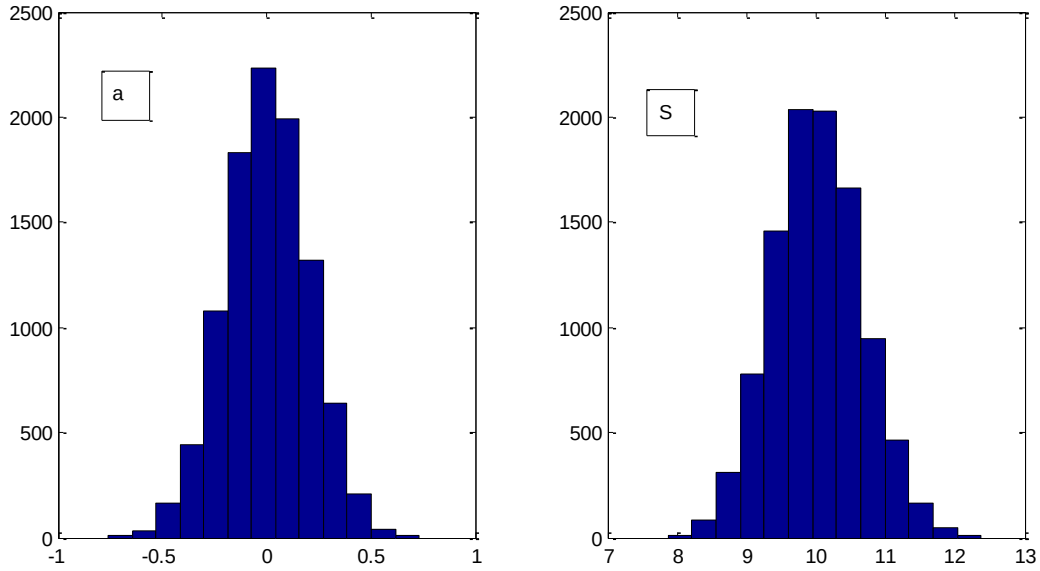


Рис. 1. Гистограммы случайных ошибок измерения длины δa и значений площади $S = ab$ прямоугольника, полученных при нормальном распределении величин a и b.

Программа затратила на все вычисления и на построение гистограмм 0.7 секунд на обычном персональном компьютере и получила следующие расчетные результаты.

$S_{\text{среднее}} = 10.0145$

$s_a = 0.2003,\ s_b = 0.1011$

Из приведенных на рисунке 1 гистограмм видно, что величины параметров модели и предсказываемой площади модели распределены по закону, очень похожему на нормальный. Численная оценка $s_S = 0.6452$, подсчитанная в программе, очень близка к теоретическому значению 0.6403, посчитанному по формуле (6). Следовательно, можно с полным основанием построить следующие утверждения:

1. Доверительный интервал $S = 10.015 \pm 0.65$ с вероятностью $Q = 68$ % покрывает истинное значение площади прямоугольника по результатам измерений.
2. Доверительный интервал $S = 10.015 \pm 2 \bullet 0.65$ с вероятностью $Q = 96$ % покрывает истинное значение площади прямоугольника по результатам измерений.
3. Доверительный интервал $S = 10.015 \pm 3 \bullet 0.65$ с вероятностью $Q = 99.7$ % покрывает истинное значение площади прямоугольника по результатам измерений.

Приведенные оценки надежностей можно в данном случае получить либо из таблиц нормального распределения, либо непосредственно из гистограммы значений величины S, приведенной на рисунке 1. В данном случае эти оценки практически совпадают. Пока мы не получили ничего неожиданного.

Проблема неопределенности закона распределения для параметров модели. Исследование на примере простой двумерной модели

Мы проверили работоспособность предложенной методики на примере нормального распределения. При другом законе распределения для входных параметров модели статистические свойства прогноза могут измениться. Потребуем от программы Rectangle, чтобы она генерировала случайные ошибки значений *a* и *b*, равномерно распределенные в пределах, соответственно, (- 0.2 + 0.2) и (- 0.1 + 0.1). Программа за 0.29 сек выдала следующие результаты.

$S_{\text{среднее}} = 10.005$

$s_a = 0.1157,\ s_b = 0.0578$

Численная оценка $s_S = 0.3699$, подсчитанная в программе, не может быть сопоставлена теоретическому значению 0.6403, посчитанному по формуле (6). Формула (6) при $\Delta a = 0.2$ и $\Delta b = 0.1$ предсказывает максимальную оценку погрешности прогноза. Но по результатам работы программы Rectangle мы получаем другую оценку максимальной погрешности. Программа находит половину разности между максимальным и минимальным случайным значением *S*. Эта оценка составляет 0.8902.

Таким образом, в случае равномерного распределения параметров *a* и *b* мы получаем иные статистические свойства прогноза. Важно, что эти свойства не согласуются с теоретическими, если под «теорией» понимать выражения (4) и (6).

Наиболее полное представление об этих свойствах дает непосредственно гистограмма случайных значений *S*. Она приведена на рисунке 2.

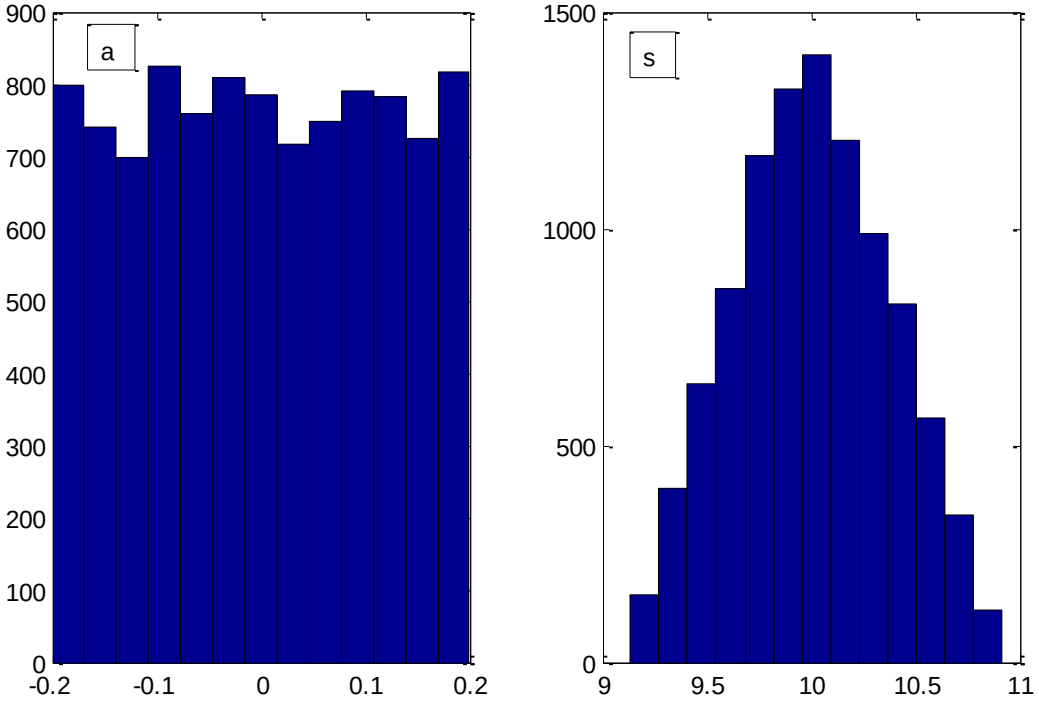


Рис. 2. Гистограммы случайных ошибок измерения длины δa и значений площади $S = ab$ прямоугольника, полученных при равномерном распределении величин a и b.

Из рисунка 2 видно, что распределение случайной величины S не подчиняется ни нормальному, ни равномерному закону.

Построим доверительные интервалы для S, опираясь непосредственно на гистограмму. Будем строить эти интервалы в единицах $s_S = 0.3699$. Получим следующие утверждения.

1. Доверительный интервал $S = 10.005 \pm 0.3699$ с вероятностью $Q = 65$ % покрывает истинное значение площади прямоугольника по результатам измерений.
2. Доверительный интервал $S = 10.005 \pm 2 \bullet 0.3699$ с вероятностью $Q = 97$ % покрывает истинное значение площади прямоугольника по результатам измерений.
3. С надежностью $Q = 100$ % возможные значения площади прямоугольника лежат в пределах $S = 10.005 \pm 0.8902$ и никогда не выходят за эти пределы.

Мы видим, что последнее утверждение заметно отличает статистические свойства S от соответствующих свойств в случае нормально распределенных параметров a и b.

Об адекватности нормального распределения природе измерений и прогнозирования

Обратим внимание на левую гистограмму на рисунке 1. Гистограмма имитирует распределение случайных ошибок измерения длины прямоугольника в выборке бутстрепа. Насколько адекватна эта гистограмма реальному процессу измерения?

Будем расширять выборку бутстрепа. В соответствии с нормальным законом распределения, гистограмма ошибок измерения будет неограниченно расширяться. Будем подсчитывать максимальный разброс значений погрешности изменения δa_{max}.

При размере выборки бутстрепа $n_b = 10000$ $\delta a_{max} = 0.7253$, что незначительно превышает величину $3s_a = 0.6$.

При размере выборки бутстрепа $n_b = 100000$ $\delta a_{max} = 0.85$, что уже превышает величину $4s_a = 0.8$.

Последний расчет потребовал 2.3 секунды машинного времени. Мощность современного персонального компьютера позволяет еще на порядок расширить выборку бутстврепа, но мы не будем этого делать. Уже ясно, что программа Rectangle со своими генераторами случайных чисел вполне правильно воспроизводит свойства нормального распределения в выборках параметров a и b модели прямоугольника. Как следствие, с расширением этих выборок расширяется и выборка предсказаний площади S. При $n_b = 10000$ $\delta S_{max} = 1.9$, а при $n_b = 100000$ $\delta S_{max} = 2.9$, что также превосходит учетверенное стандартное отклонение предсказываемой величины площади S.

Итак, если считать нормальное распределение универсальным законом, адекватным природе измерений и моделирования, то мы получим неизбежное следствие в виде утверждения.

При прогнозировании явлений на основе моделей с нормально распределенными параметрами можно ожидать, пусть с очень малой вероятностью, любых отклонений прогноза от истинного значения прогнозируемой характеристики изучаемого явления.

Практика измерений с помощью современных физических приборов убедительно показывает, что погрешности измерений всегда ограничены. Следовательно, этот эмпирический факт надо как-то учитывать в практике моделирования при оценке ошибок прогноза. Это отмечено в работе [2], где предлагается процедура построения ограниченных интервальных оценок точности прогноза при работе с формальными статистическими моделями многомерных данных сложной природы. В данной вычислительной процедуре, предназначенной работе с содержательными моделями, предлагается явно учитывать факт ограниченности погрешностей входных параметров при получении имитационных выборок прогнозируемых характеристик явления. Проверим сначала работоспособность этого предложения на простом примере модели площади прямоугольника.

Впрочем, одну такую проверку мы уже выполнили, когда в программе Rectangle задали равномерный закон распределения для параметров a и b. При $n_b = 10000$ оценка максимального разброса значений S составляет $\delta S_{max} = 0.8902$. При $n = 100000$ $\delta S_{max} = 0.898$. Эти оценки можно считать неотличимыми, поскольку δS_{max} является случайной величиной.

В программе Rectangle есть возможность задать закон распределения, по форме похожий на нормальный, но с ограниченной областью определения случайной величины. Ограничение задается как требование, чтобы случайная величина не выходила за пределы $(- s_{limit}*\sigma, s_{limit}*\sigma)$. Зададим $s_{limit} = 2$, $n_b = 10000$, и получим следующие результаты.

$S_{среднее} = 9.996$

$s_a = 0.176$, $s_b = 0.09$

$s_S = 0.569$, $\delta S_{max} = 1.7$

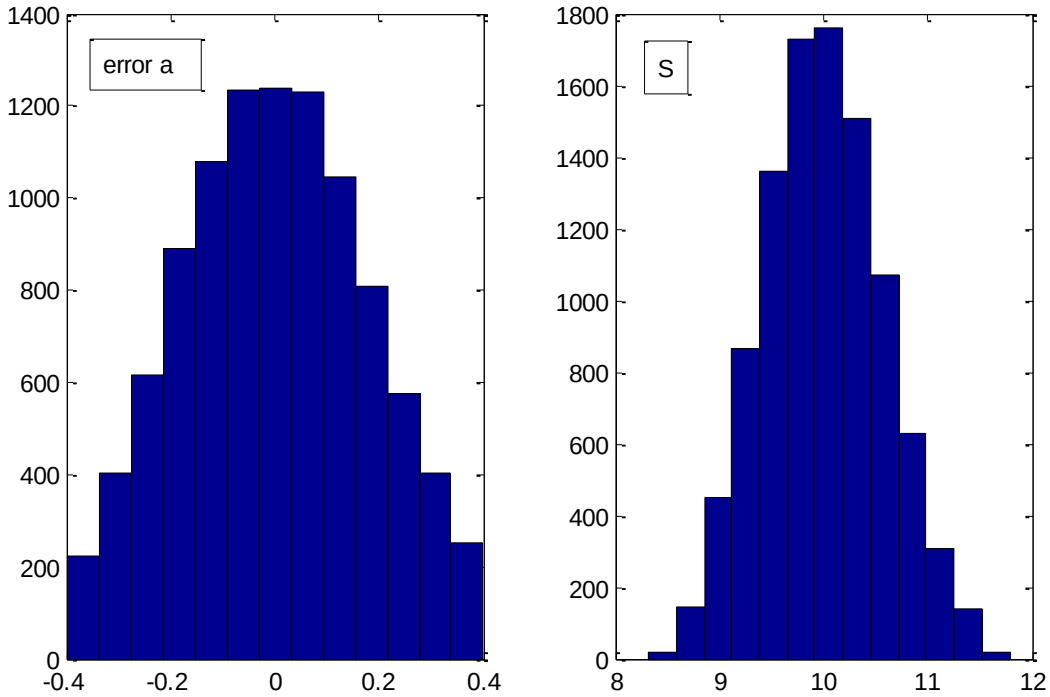


Рис. 3. Гистограммы случайных ошибок измерения длины δa и значений площади $S = ab$ прямоугольника, полученных при ограниченном нормальном распределении величин a и b.

Из рисунка 3 видно, что закон распределения ошибок величины a отличается от нормального отсутствием крыльев, но прогнозируемая величина S распределена по закону, напоминающему нормальный. Из приведенных численных данных видно, что ограниченность распределения параметров модели привела к меньшему разбросу величины S. Однако основная особенность данного имитационного вычислительного эксперимента проявляется при расширении объема выборки бутстрепа. При $n_b = 100000$ все статистические свойства величины S остаются такими же, как при $n = 10000$.

Следовательно, и величина S распределена по закону, напоминающему нормальный закон с ограниченной областью определения.

Итак, если считать ограниченность ошибок универсальным законом, адекватным природе измерений, то мы получим неизбежное следствие в виде значительно более оптимистичного утверждения.

> **При прогнозировании явлений на основе моделей с полученными в эксперименте параметрами можно уверенно ожидать, что прогноз даст лишь ограниченные отклонения от истинного значения прогнозируемой характеристики изучаемого явления.**

Данное утверждение представляется очень важным для всей практики моделирования с целью разработки аналитических методик. Конечно, требуется еще прояснить условия,

когда оно применимо, а также требуется обосновать закон ограниченности погрешностей измерений. Однако это предполагает проведение специального исследования. Теперь же мы перейдем к проверке работоспособности предлагаемой методики в случае прогнозирования колебательного спектра молекулы. Закончим данную главу демонстрацией применения предложенной процедуры оценки точности теоретического спектра при разработке методики безэталонного спектрального анализа.

4.2. Статистические свойства теоретического колебательного спектра молекулы

В системе LevInfinite имеются готовые модели этана и этилена, находящиеся в папках EthaneML и EthyleneML соответственно. Обе модели полностью воспроизводят материал, ранее опубликованный в нашей книге [3]. Это головные модели соответствующих гомологических рядов соединений – алканов и олефинов. Силовые и электрооптические параметры моделей получены в процессе постановки и решения обратных спектральных задач. Это накладывает некоторые особенности на статистические свойства приведенных в книге параметров.

Дело в нашей постановке обратных спектральных задач. Мы следовали логике химиков, которые мыслят свойства не отдельных соединений, а целых гомологических рядов соединений, поскольку в этих рядах проявляются некоторые устойчивые признаки функциональных групп, входящих во все молекулы данного ряда. В ряду молекул повторяются как химические, так и физические свойства. Следуя этому эмпирическому закону, мы при решении обратных задач требовали, чтобы параметры, найденные для головных молекул ряда, хорошо воспроизводили спектры всех молекул ряда, даже тех, для которых обратные задачи не решались. Такое свойство параметров называется их переносимостью в ряду родственных моделей. Более того, мы требовали, чтобы параметры, найденные для головных молекул ряда, хорошо воспроизводили спектры молекул других рядов, куда входят структурные группировки данного ряда. Именно такая постановка обратных задач дала нам возможность собрать сравнительно небольшую коллекцию так называемых стандартных колебательных моделей молекул, пригодную для прогнозирования неопределенно широкой совокупности сложных молекул с фрагментами стандартных моделей.

За такое удовольствие приходится чем-то расплачиваться. Не только трудом и расходом машинного времени, но и свойствами найденных параметров. При подобной постановке обратных задач невозможно получить такие параметры, чтобы модели прогнозировали спектры, совпадающие с экспериментальными в пределах точности спектральных приборов. Обязательно остаются неустранимые невязки частот и интенсивностей поглощения в ИК спектрах. Мы использовали метод наименьших квадратов, который позволяет свести к минимуму сумму квадратов невязок, остающихся после решения обратных задач. В теорию метода наименьших квадратов органически входит процедура оценки статистических свойств решения задачи. Оказывается, что можно не анализировать причины неустранимости невязок, а считать их результатами случайных ошибок измерения определяемых параметров. Это, конечно, не так. Мы понимаем, что

виноваты недостатки самих моделей и теории, а также требование переносимости параметров. Конечно, в конкретных молекулах данного ряда параметры могут и должны изменяться по мере усложнения структуры соединений. Но все эти причины можно на модельном же уровне считать случайными. Тогда мы приписываем и невязкам, и ошибкам параметров свойства случайных величин, распределенных по нормальному закону с индивидуальными дисперсиями.

Итак, метод наименьших квадратов дает возможность одновременно определить параметры модели по экспериментальным спектрам и оценить средние статистические отклонения этих параметров. В среднем эти отклонения не выходили за пределы 5% от значения самого параметра. Сейчас перед нами стоит задача проследить, как статистические свойства параметров модели сказываются на статистических свойствах прогноза, даваемого расчетом колебательного спектра на основе данной модели и данной теории. Конечно, мы определим не полные погрешности частот и интенсивностей в теоретическом спектре, а только статистическую составляющую этих погрешностей. Это уже кое-что, поскольку без таких оценок погрешностей невозможно разрабатывать какие-либо аналитические методики, опирающиеся на спектроскопические эксперименты.

Модель этана

Вот как эта модель выглядит в окне программы Umiu.m:

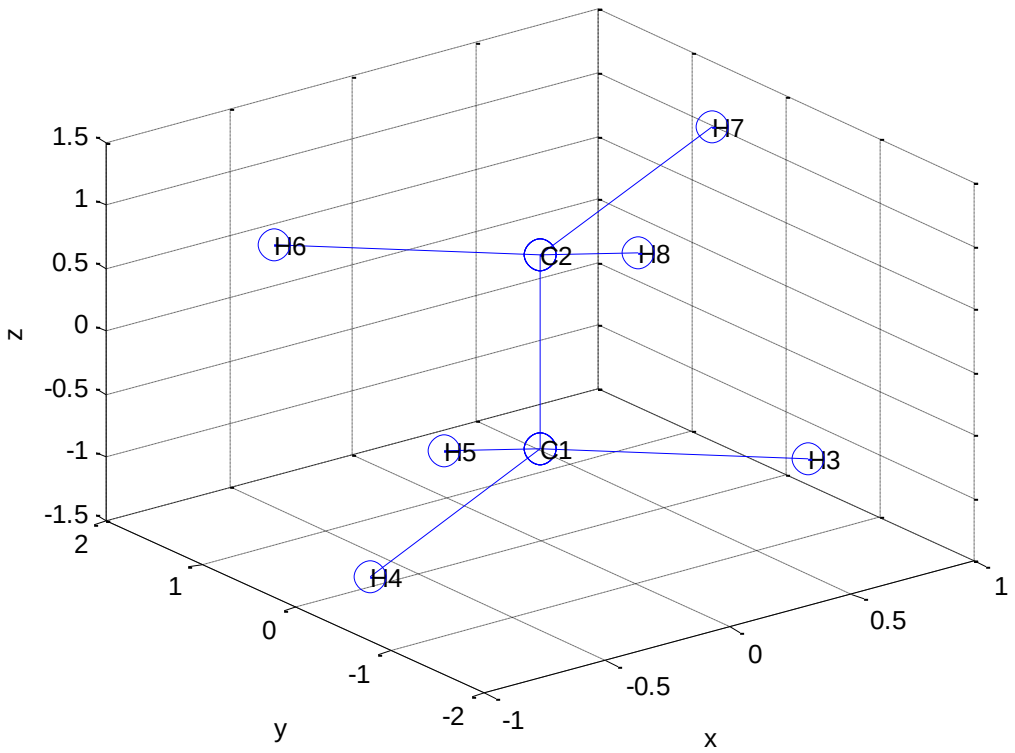


Рис. 5. Модель молекулы этана.

Характерные силовые и электрооптические параметры модели в файле u_list.txt

1 1 6.74 0.0 0.0 0.0 qq 1 2
1 8 0.46 NaN 0.0 NaN qa 1 2 / 2 1 3
2 2 8.03 0.305 0.79 0.0 qq 1 3
2 3 0.06 NaN 0.28 NaN qq 1 3 / 1 4
2 8 0.3 NaN -0.4 NaN qa 1 3 / 2 1 3
2 9 0.0 NaN -0.535 NaN qa 1 3 / 2 1 4
2 14 0.3 NaN -0.3 NaN qa 1 3 / 3 1 4
2 16 0.0 NaN -0.38 NaN qa 1 3 / 4 1 5
8 8 0.92 NaN NaN NaN aa 2 1 3
8 9 -0.025 NaN NaN NaN aa 2 1 3 / 2 1 4
8 11 0.14 NaN NaN NaN aa 2 1 3 / 1 2 6
8 12 -0.02 NaN NaN NaN aa 2 1 3 / 1 2 7
8 14 -0.034 0 NaN NaN NaN aa 2 1 3 / 3 1 4
14 14 0.71 NaN NaN NaN aa 3 1 4
14 15 -0.034 NaN NaN NaN aa 3 1 4 / 3 1 5

В каждой строке:

1. Номер *i* первой колебательной координаты
2. Номер *j* второй колебательной координаты
3. Силовая постоянная взаимодействия координат *i* и *j*
4. Дипольный момент связи. Если это не связь или для связи $i \neq j$, то стоит NaN – Not a Number; это значение не учитывается в расчете
5. Производная дипольного момента данной связи номер *i* по колебательной координате номер *j*. Если это не связь, то стоит NaN
6. Производная дипольного момента связи номер *j* по колебательной координате номер *i*. Если *i* и *j* не связи или связи эквивалентны, то стоит NaN
7. К какому типу взаимодействующих колебательных координат относится строка
8. Номера атомов во взаимодействии координат через слэш

Модель с этими параметрами и теория дают прогноз частот и ИК интенсивностей, а мы навязываем полосам в спектре еще и полуширины, похожие на экспериментальные:

978.68 0 10
1381.83 0 10
2896.22 0 10
1376.63 1.88 8
2891.65 25.57 30
820.87 0.70 30
1462.53 1.70 10
2973.87 16.28 30
820.87 0.70 30
1462.53 1.70 10
2973.87 16.28 30
1182.16 0 10
1455.14 0 10
2963.10 0 10
1182.09 0 10
1455.14 0 10
2963.08 0 10

Там, где интенсивность нулевая, полуширина не имеет значения.

Получаем прогноз спектральной кривой:

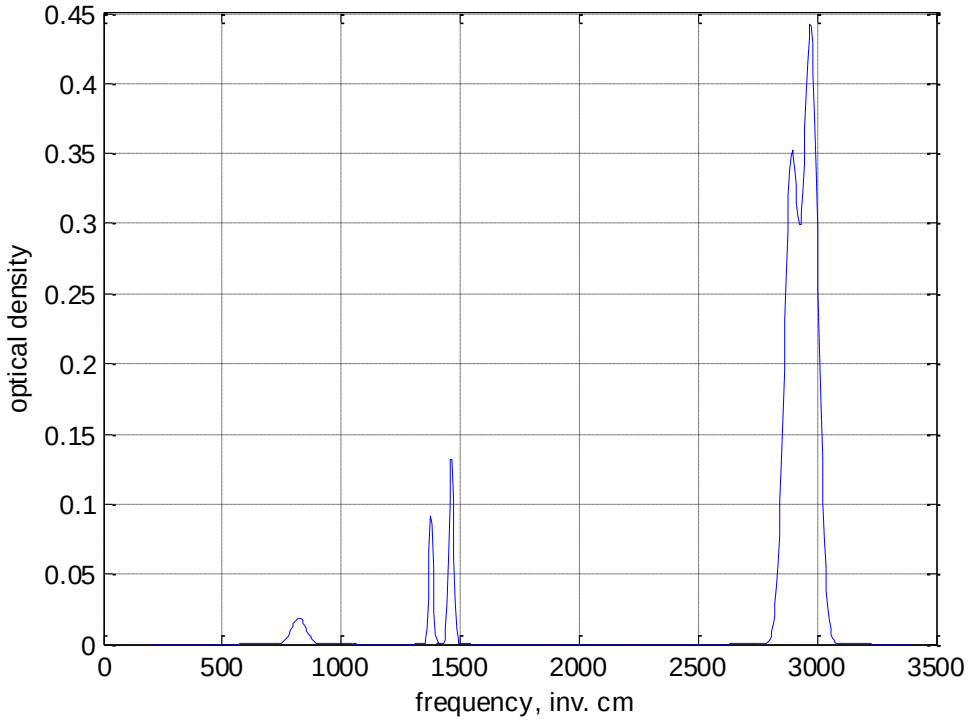


Рис. 6. Теоретический спектр ИК поглощения молекулы этана

Ординаты этой спектральной кривой запоминаются в файле IRcurve0.txt, чтобы в имитационном эксперименте по безэталонному спектральному анализу кривая представляла собой «приборный спектр чистого вещества». Мы будем вносить этот спектр, умноженный на соответствующую концентрацию, в суммарную «экспериментальную» кривую.

Задаем, какие параметры варьировать и по какому закону. Это можно сделать с помощью утилиты set_params_std.m. Утилита просматривает файл Umiu.m из папки данной модели (имя папки должно содержаться в файле ToCalculate.ini). Для любого параметра, если он не равен нулю и не NaN, утилита назначает предел варьирования, равный 5% от абсолютного значения параметра. В папку модели утилита помещает текстовой файл params_std.txt, структура которого соответствует структуре файла u_list.txt. Поэтому при желании можно приготовить файл params_std.txt вручную, взяв за основу содержимое файла u_list.txt. Однако это опасно, поскольку количество и порядок строк в файле params_std.txt должно строго соответствовать структуре файла Umiu.m. При ручной подготовке тут можно легко ошибиться, а утилита данное соответствие обеспечивает автоматически.

Вот как выглядит файл params_std.txt для модели этана.

```
0.33700   0.00000   0.00000   0.00000
0.02300   0.00000   0.00000   0.00000
0.40150   0.01525   0.03950   0.00000
0.00300   0.00000   0.01400   0.00000
0.01500   0.00000   0.02000   0.00000
0.00000   0.00000   0.02675   0.00000
```

```
0.01500   0.00000   0.01500   0.00000
0.00000   0.00000   0.01900   0.00000
0.04600   0.00000   0.00000   0.00000
0.00125   0.00000   0.00000   0.00000
0.00700   0.00000   0.00000   0.00000
0.00100   0.00000   0.00000   0.00000
0.00170   0.00000   0.00000   0.00000
0.03550   0.00000   0.00000   0.00000
0.00170   0.00000   0.00000   0.00000
```

Первый столбец относится к данным о варьировании силовых постоянных, второй – к дипольным моментам, третий и четвертый – к производным от дипольных моментов по колебательным координатам. Там, где стоит ноль, параметр варьировать не надо. Если сравнить эту таблицу с таблицей значений параметров, то ясно, что предлагается варьировать одновременно все ненулевые параметры модели.

Этот файл является файлом исходных данных для программы IRSerrors.m, которая будет варьировать параметры модели и выдавать информацию о статистике частот и интенсивностей.

Теперь о статистическом законе, которым будет руководствоваться IRSerrors.m, создавая различные сочетания параметров в каждом обороте своего главного цикла. IRSerrors.m берет исходное значение параметра и прибавляет случайное число, распределенное по нормальному закону с ограничениями. Берем случайное число x у генератора, имитирующего закон Гаусса с параметрами 0 и σ. Если x выходит за пределы $-limit*\sigma < x < limit*\sigma$, то такое число мы не принимаем. Если x входит в эти пределы, то принимаем за добавку к параметру модели.

В таблице данных для программы IRSerrors, показанной выше, приведены значения σ для каждого варьируемого параметра. В самой программе IRSerrors задано значение $limit = 1.0$. Это имитирует реальную ситуацию с немногочисленными измерениями, когда ошибки измерений редко выходят за пределы ±σ (вероятность наблюдать погрешность в таких пределах равна 68 %).

В программе также задано число попыток прогнозирования, которые должен выполнить бутстреп. В данном эксперименте это 200. Мой опыт работы с бутстрепом показывает, что этого достаточно для получения представительных статистических данных о случайном явлении, если оно не слишком экзотично. Наша теория – не экзотика.

Программа IRSerrors проработала 34 секунды и дала следующие результаты.

```
 978.91 13.89  0.00  0.00
1381.22 14.06  0.00  0.00
2896.77 36.13  0.00  0.00
1376.00 13.64  1.92  0.52
2892.19 36.20 25.62  1.01
 821.19 12.41  0.74  0.23
1460.26 16.64  1.74  0.42
2974.51 38.16 16.36  1.38
 821.14 12.40  0.74  0.23
```

```
1460.27 16.64  1.74  0.42
2974.49 38.16 16.36  1.38
1182.01 11.59  0.00  0.00
1453.30 15.27  0.00  0.00
2963.73 38.33  0.00  0.00
1181.94 11.59  0.00  0.00
1453.30 15.27  0.00  0.00
2963.71 38.33  0.00  0.00
```

В каждой строчке даны: среднее значение частоты из 200 прогнозов бутстрепа, среднее квадратичное отклонение частоты, средняя абсолютная интенсивность поглощения, среднее квадратичное отклонение интенсивности.
Эли данные сохраняются в файле IRSstatistics.txt. Индивидуальные значения 200 пар случайных частот и интенсивностей, полученные в цикле бутстрепа, сохраняются в файле FreqsInts.mat для дальнейшего использования в имитационном эксперименте по разработке методики безэталонного спектрального арализа.

Для каждого набора параметров модели, создаваемого бутстрепом, рассчитана полная спектральная кривая. Все эти кривые наложены друг на друга:

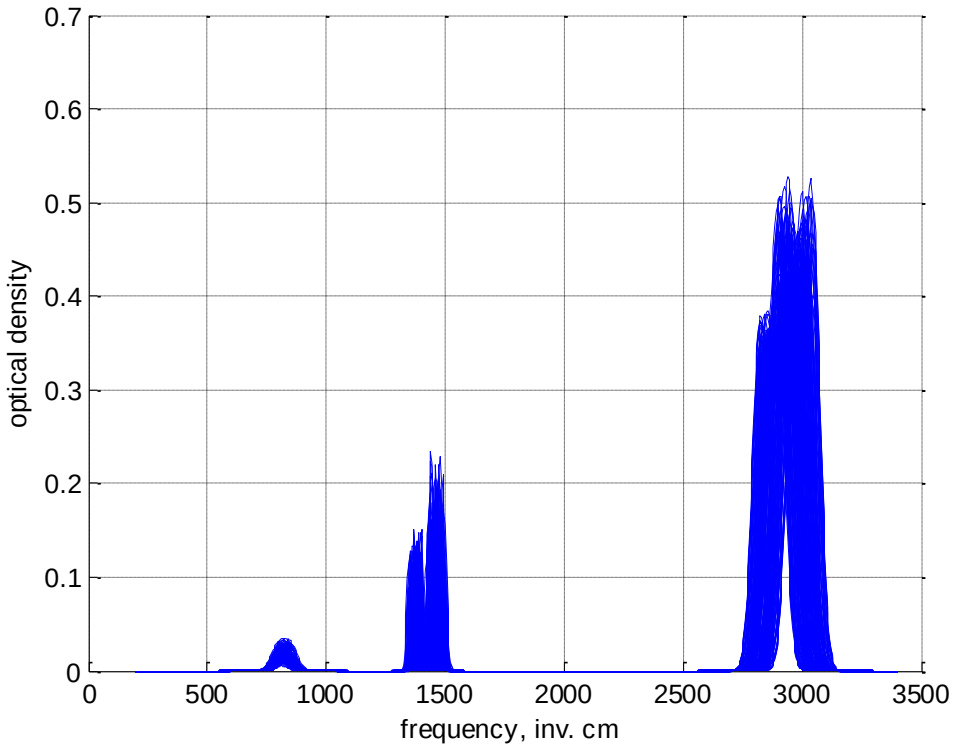


Рис. 7. Спектральные кривые ИК поглощения модели этана при вариации параметров модели в 200 попытках бутстрепа.

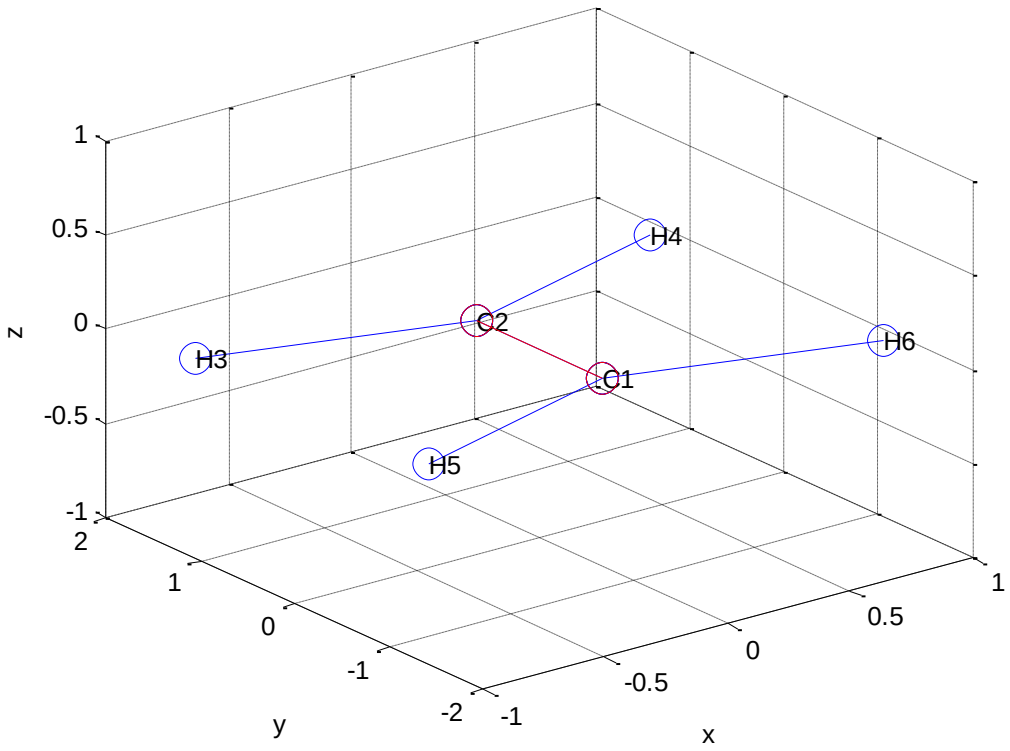
На рисунке 7 видно, что неопределенность в значениях параметров модели этана приводит к заметной неопределенности в положениях и интенсивностях полос ИК поглощения. Причем, вся качественная картина статистических свойств нашего прогноза видна просто с первого взгляда. А из таблицы средних частот и их среднеквадратичных отклонений видно, что ошибки параметров порядка 5% от значений самих параметров привели не 5% дополнительного уширения спектральных полос поглощения. Скорее, это 10%. С абсолютными интенсивностями в некоторых полосах дело обстоит еще хуже.

Заметим, что предложенная процедура и программа IRSerrors могут найти самые различные применения в статистической физике. Вспомните, с каким трудом выводятся длинные формулы, когда надо усреднять микроскопические характеристики ансамбля молекул, чтобы добраться до характеристик, доступных макроскопическим приборам. А можно обойтись безо всяких формул. Представим себе, что надо решить следующую задачу.

Некое органическое вещество находится в нейтральном растворителе, который не дает собственного ИК поглощения. Надо разобраться в причинах уширения полос поглощения исследуемого органического вещества. Участвуя в тепловом движении, молекулы вещества подвергаются атакам, как со стороны молекул растворителя, так и друг друга. Слабые электрические взаимодействия оказывают влияние на внутреннюю электрическую жизнь молекул вещества, возмущая и временно изменяя значения его параметров. Это случайные процессы. Сделав какие-то предположения об этих влияниях, обратимся к программе IRSerrors и приготовим для нее исходные данные, которые будут имитировать изменения избранных параметров модели вещества в заданных пределах. Мы получим картину, похожую на рисунок 7. Если согласие с экспериментом будет неудовлетворительным, то придется пересмотреть наши представления о механизмах взаимодействия молекул вещества друг с другом и с растворителем. В конце концов, есть надежда угадать правильно и получить от численного эксперимента подтверждение, что мы что-то в этом понимаем.

Модель этилена

Вот как эта модель выглядит в окне программы Umiu.m:



Силовые и электрооптические параметры модели

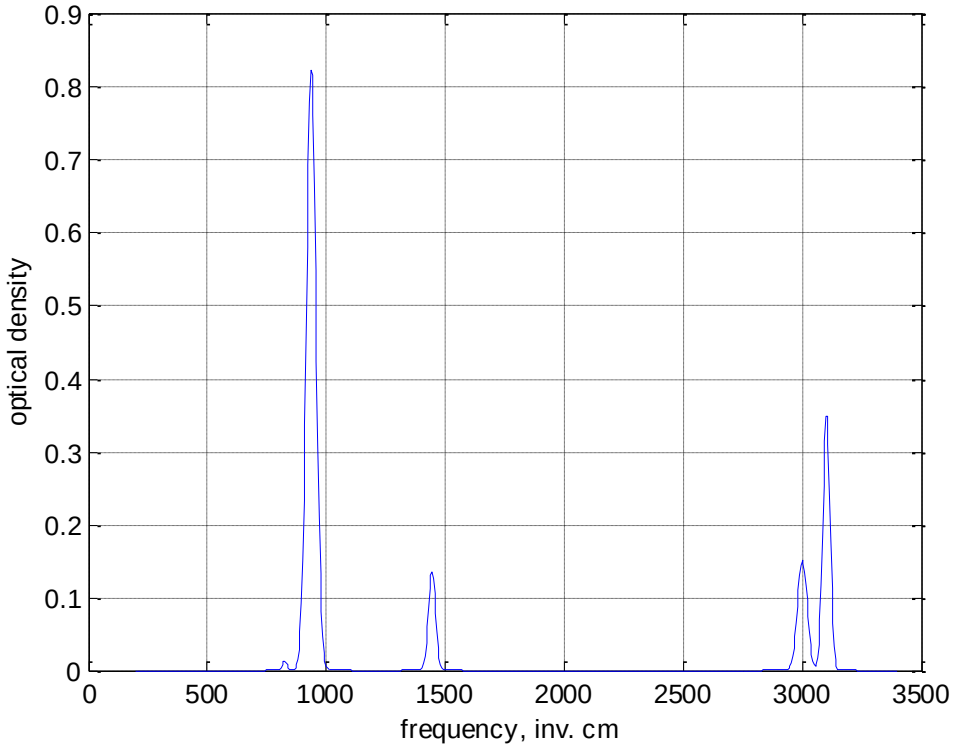
```
  1    1   14.2    0.0    0.0    0.0 qq     1    2
  1    2   0.1     NaN    0.0    NaN qq     1    2 /     2    3
  1    6   0.42    NaN    0.0    NaN  qa     1    2 /     1    2    3
  2    2   8.55    0.719  0.493  0.0   qq     2    3
  2    3   0.030   NaN   0.041   NaN  qq     2    3 /     2    4
  2    6   0.2     NaN   0.481   NaN  qa     2    3 /     1    2    3
  2   10   0.12    NaN   0.551   NaN   qa     2    3 /     3    2    4
  6    6   0.77    NaN   NaN     NaN   aa     1    2    3
  6    8   -0.025  NaN   NaN     NaN   aa     1    2    3 / 2  1    5
  6    9   0.095   NaN   NaN     NaN   aa     1    2    3 / 2  1    6
 10   10   0.57    NaN   NaN     NaN   aa     3    2    4
 12   12   0.35    NaN   NaN     NaN   rr     3    2    4    1
 12   13   0.047   NaN   NaN     NaN   rr     3    2 4 1 / 6 1 5 2
 14   14   0.755   NaN   NaN     NaN   hh     3    2  4  5   1    6
```

Модель с этими параметрами дает прогноз частот и ИК интенсивностей, а мы навязываем полосам в спектре еще и полуширины, похожие на экспериментальные:

```
1339.8   0.0   10
1620.9   0.0   10
3017.4   0.0   10
1246.8   0.0   10
3099.1   0.0   10
824.6    0.32  10
3102.2   13.3  15
1444.4   5.1   15
3000.9   7.53  20
936.8    41.42  20
957.1     0.0  10
1020.1    0.0  10
```

Там, где интенсивность нулевая, полуширина не имеет значения.

Получаем прогноз спектральной кривой:



Эта спектральная кривая запоминается, чтобы в имитационном эксперименте для публикации она представляла собой «приборный спектр чистого вещества». Мы будем вносить этот спектр, умноженный на соответствующую концентрацию, в суммарную «экспериментальную» кривую.

Задаем, какие параметры варьировать и по какому закону. Так выглядит файл исходных данных для программы IRSerrors:

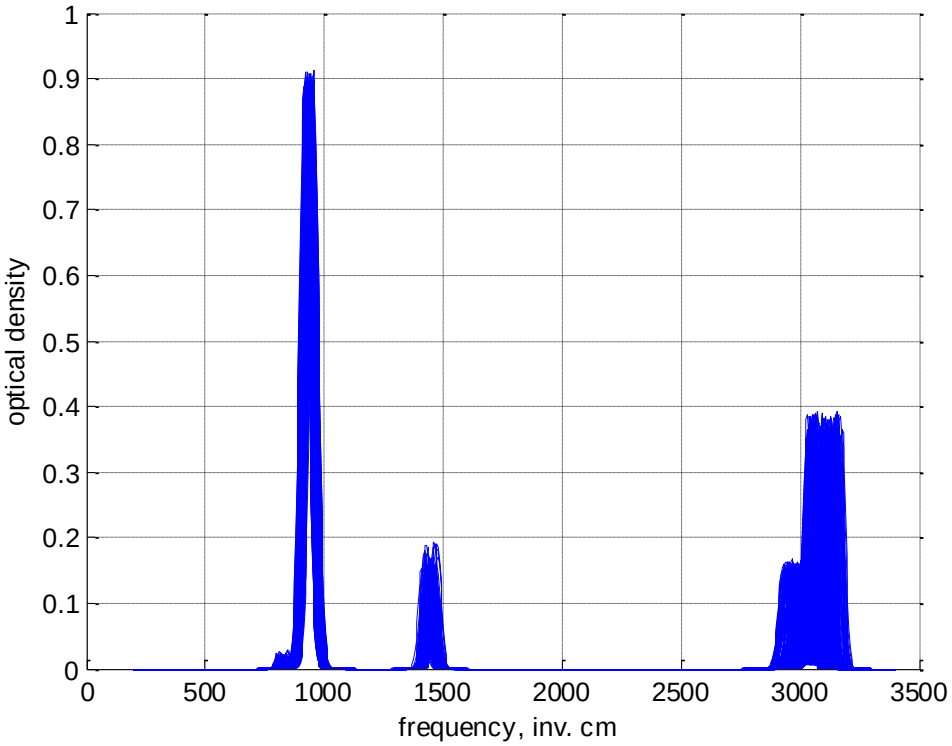
```
0.71000   0.00000   0.00000   0.00000
0.00500   0.00000   0.00000   0.00000
0.02100   0.00000   0.00000   0.00000
0.42750   0.03595   0.02465   0.00000
0.00150   0.00000   0.00205   0.00000
0.01000   0.00000   0.02405   0.00000
0.00600   0.00000   0.02755   0.00000
0.03850   0.00000   0.00000   0.00000
0.00125   0.00000   0.00000   0.00000
0.00475   0.00000   0.00000   0.00000
0.02850   0.00000   0.00000   0.00000
0.01750   0.00000   0.00000   0.00000
0.00235   0.00000   0.00000   0.00000
0.03775   0.00000   0.00000   0.00000
```

Программа IRSerrors проработала 24 секунды и дала следующие результаты.

```
1340.16        8.09     0.00     0.00
```

1624.07	16.48	0.00	0.00
3015.83	34.75	0.00	0.00
1247.76	11.74	0.00	0.00
3097.36	36.68	0.00	0.00
826.31	10.37	0.09	0.06
3100.42	36.70	13.53	0.56
1444.65	9.66	7.45	0.99
2999.14	35.17	7.82	0.24
936.88	6.68	49.34	2.03
957.21	8.98	0.00	0.00
1018.73	10.20	0.00	0.00

Для каждого набора параметров модели, создаваемого бутстрепом, рассчитана полная спектральная кривая. Все эти кривые наложены друг на друга:



4.3. Имитация процедуры безэталонного анализа для смеси этана и этилена

Идея безэталонного анализа заключается в следующем. Количественный анализ смеси органических веществ становится возможным без предварительной подготовки материальных эталонов, когда зарегистрированный спектральным прибором сигнал от смеси сравнивается с теоретически предсказанными спектрами смеси при варьировании предполагаемых концентраций компонент смеси [4]. Успех и точность количественного анализа при этом определяется как погрешностями экспериментального спектра смеси, так и точностью прогнозирования теоретических спектров. К сожалению, при написании книги [4] теория погрешностей прогноза оптических спектров сложных органических соединений еще не была разработана, и поэтому проблема априорной оценки точности безэталонного спектрального анализа даже не ставилась. Теперь, как мы видели, предложены пути решения указанной проблемы. Это оказало влияние на саму постановку

проблемы проведения количественного спектрального анализа смеси веществ без предварительной подготовки материальных эталонов для анализируемых смесей. Поэтому следует пояснить, как теперь автор идеи безэталонного анализа, Л.А. Грибов ставит задачу, прямо опираясь на описанную выше процедуру оценки точности прогноза спектральной кривой ИК поглощения смеси веществ. Вот как выглядит предложение Л.А. Грибова.

Пусть в эксперименте получена оцифрованная спектральная кривая для смеси веществ. Ординаты кривой $y(\nu)$ выражены в единицах оптической плотности, точки ν на оси абсцисс выражены в $см^{-1}$. Временно будем считать, что $y(\nu)$ не содержит экспериментальных ошибок. Предполагается также, что дискретизация при переводе аналоговой спектральной кривой в цифровую форму проведена без потерь быстрых изменений в спектре. То есть, точки на оси абсцисс расположены достаточно густо.

Пусть мы располагаем надежной информацией или обоснованным предположением, какие органические вещества входят в исследуемую смесь. Следовательно, задача будет состоять лишь в определении вектора c концентраций компонент смеси.

Пусть мы располагаем спектральными кривыми $Y(\nu)$ компонент смеси. Кривые $Y(\nu)$ представлены в той же форме, что и $y(\nu)$, то есть точки ν на оси частот у всех кривых одни и те же. Ординаты $Y(\nu)$ соответствуют единичным концентрациям веществ, например, даны на одну молекулу вещества, либо на 1 ppm, либо на другую удобную в данном эксперименте единицу.

Тогда имеет место тождество

$$y = Yc. \tag{7}$$

Здесь y – матрица-столбец значений оптических плотностей в $y(\nu)$; Y – прямоугольная матрица, где по столбцам записаны оптические плотности в спектрах $Y(\nu)$ компонент смеси; c – матрица-столбец концентраций веществ в данной смеси.

Умножим обе части равенства слева на псевдообратную матрицу Y^{-1}. Получим

$$Y^{-1}y = c. \tag{8}$$

Мы получили матричное уравнение для определения концентраций. В случае идеальных $Y(\nu)$ и $y(\nu)$ уравнение (8) даст точные концентрации компонент смеси. Если же $Y(\nu)$ и $y(\nu)$ содержат погрешности, то мы получим приближенное выражение для вектора c.

Идея безэталонного анализа состоит в использовании не экспериментальных спектров высокой точности $Y(\nu)$, а теоретических. Если бы мы всегда располагали точными экспериментальными спектрами определяемых веществ, то теория ИК спектров молекул была бы совершенно не нужна при разработке методики такого спектрального анализа. Но невозможно заранее располагать экспериментальными спектрами любых органических веществ, которые будут только завтра синтезированы или извлечены из природных объектов. А вот разумные предположения об их структурах химик может и должен сделать. Тогда, пользуясь стандартными моделями с переносимыми параметрами, исследователь сконструирует правдоподобную модель каждого нового предполагаемого соединения и получит прогноз их теоретических спектров. Но мы видели, что теоретические колебательные спектры отягчены погрешностями, полученными в наследство от параметров моделей. Возникает необходимость, заменив в уравнении (7) Y

на матрицу теоретических спектров, вычислить *c* и найти погрешности всех найденных значений компонент этого вектора. Здесь Л.А. Грибов предлагает такой вычислительный алгоритм.

Будем выполнять цикл бутстрепа, на каждом шаге которого в матрицу *Y* подставлять случайные экземпляры теоретических спектров, полученные предварительно при варьировании параметров моделей описанным ранее способом. Для данного шага составим и решим уравнение (8). Накопим все случайные экземпляры вектора *c*, полученные в цикле бутстрепа. Это даст возможность найти статистические свойства вектора концентраций непосредственно из гистограмм, построенных для компонент вектора.
Проверим работоспособность предлагаемого алгоритма с помощью имитационного компьютерного эксперимента.

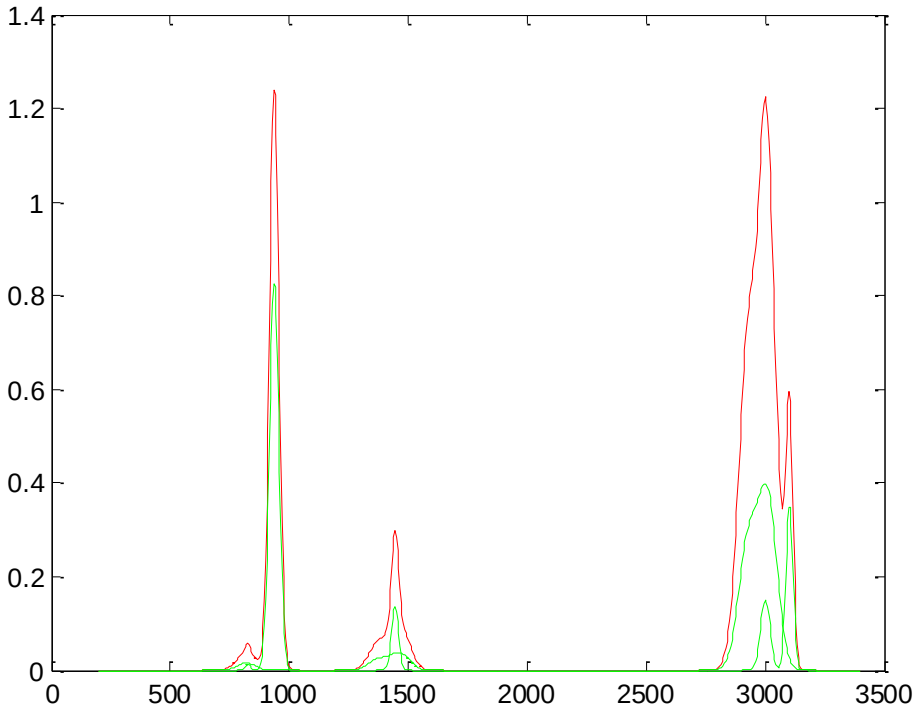
Вместо экспериментального ИК спектра смеси приготавливаем сумму стандартных спектральных кривых с помощью специально написанной программы pre_ne.m. Эта программа занимается предварительным (pre) сбором данных для проведения безэталонного (Non-Etalon = ne) анализа смеси из *n* компонент. В нашем случае *n* = 2. Для программы pre_ne.m и для следующих процедур анализа надо заготовить в папке Config файл nonetalon_n.txt следующего содержания

2 число компонент в смеси
EthaneML
2.5
EthyleneML
1.5

После каждого имени папки с моделью, включаемой в анализ, указана та концентрация, с которой ИК спектр данной модели входит в «экспериментальный» спектр смеси. Концентрации указаны в молекулах, попавших в оптический путь спектрального прибора, поскольку в наших теоретических спектрах интенсивность дается на одну молекулу.

Перед запуском программы pre_ne.m рекомендуется очистить рабочую область МатЛаб, поскольку программы безэталонного анализа оформлены как скрипты, а не как функции. В связи с этим рабочая область МатЛаб будет открыта для обозрения, и не хотелось бы, чтобы к нужным переменным примешивались остатки предыдущих расчетов.

Программа pre_ne.m выводит на экран и запоминает такую «экспериментальную» спектральную кривую смеси



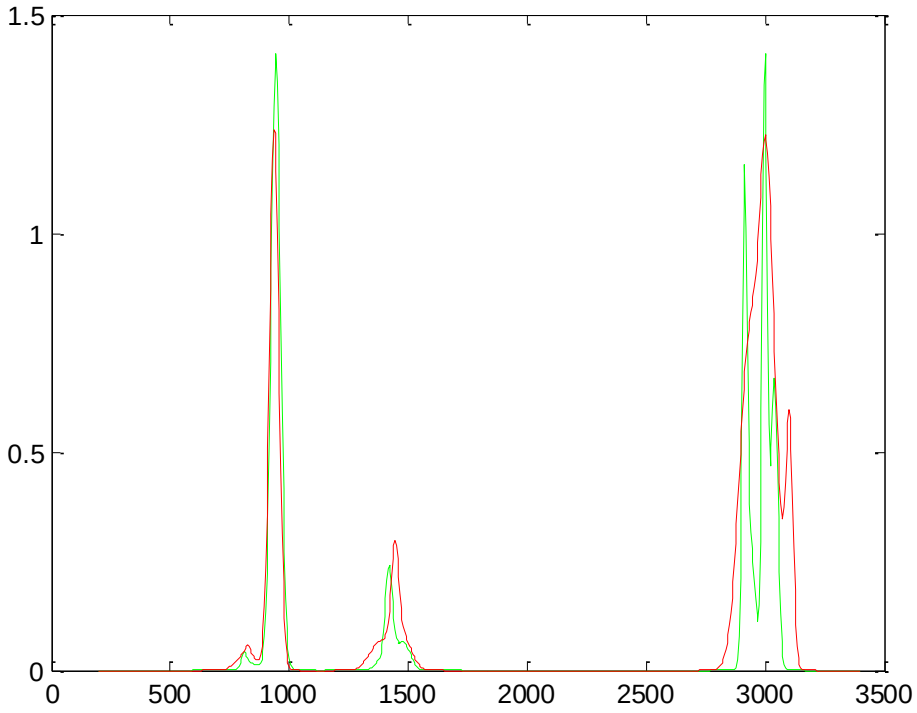
«Экспериментальный» спектр смеси показан красным цветом, а входящие в смесь компоненты , в расчете на одну молекулу, – зеленым.

Далее запускаем на счет программу безэталонного анализа nonetalon_n.m. Результаты работы программы находим в рабочей области.

i_best = 65 эту переменную я объясню потом
discrepancy = 11.4
c_bestB = [1.02, 1.79] вместо заданных концентраций [2.5, 1.5]
c1c2 содержит информацию о статистике концентраций в решениях бутстрепа:

	Mean	Std	min	max
$C_{этан}$	0,80	0,23	0,34	1,26
$C_{этилен}$	1,56	0,19	1,05	1,92

Сравним теоретическую кривую, которую предсказывает программа безэталонного анализа, складывая спектры чистых компонент, умноженные на найденные значения концентраций.



«Экспериментальная» кривая дана красным цветом, теоретическая – зеленым.

Даже не очень внимательный читатель заметит, что полученные результаты полностью дискредитируют идею Грибова.

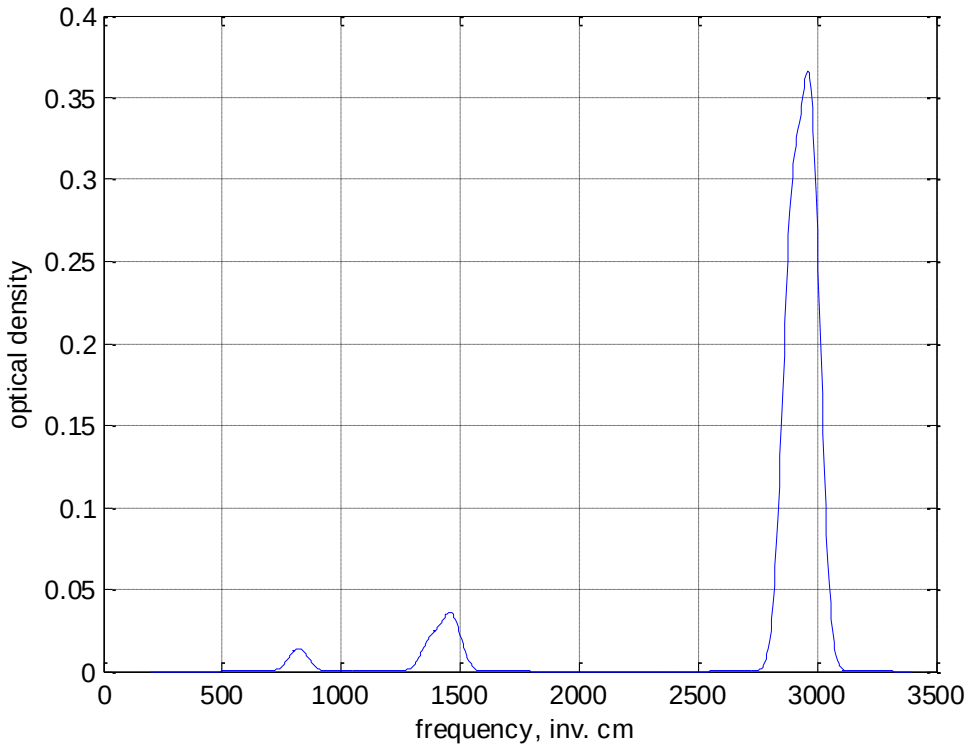
Вопрос – зачем же я привожу такие результаты? В научной литературе, вроде бы, не принято приводить результаты, которые гробят собственную идею. Но это не совсем обычная книга. Слава Интернету – он позволяет делать и то, что не принято. Я хочу на данном материале показать – путь от теории к возможности решать прикладные задачи, в норме, тернист. Обычно об тернии сдирают кожу инженеры, берущиеся за продвижение теории в приложения. Наш случай интересен тем, что гипотетические инженеры должны были бы обладать исключительно хорошей физической базой знаний и умений. Чего наша система подготовки инженеров на предполагала. Отсюда в Советское время следовал вполне предсказуемый результат: за все время развития вычислительной техники советские инженеры не предложили ни одного нового материального элемента для компьютеров.

Итак, мы не могли полагаться на улучшение наших результатов с помощью инженеров. Обдирать кожу об тернии пришлось самому Грибову. Он, будучи глубоким физиком, рассудил так.

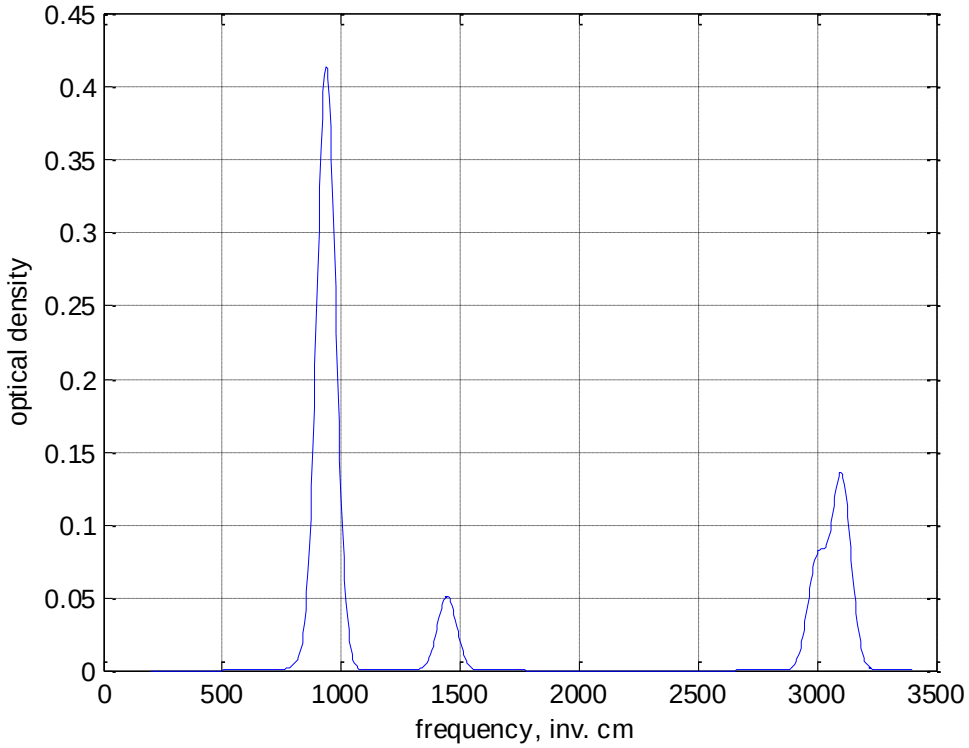
Мы, затевая безэталонный анализ, не знаем экспериментального спектра ни одного чистого компонента смеси. Поэтому надеяться на совпадение по форме нашего модельного спектра с реальным спектром в сумме не приходится. Полосы поглощения модели даже для одного чистого вещества будут сдвинуты относительно полос поглощения в реальном спектре. Из теории анализа видно, что в решении задачи концентрация данного компонента будет заниженной. Если острые полосы совсем не перекроются, то результат будет даже нулевым. Значит, надо учесть физическую реальность ИК спектроскопии. А она состоит в том, что в эксперименте полосы поглощения, имеющие очень малую естественную ширину, проходят через прибор с довольно

широкой аппаратной функцией. Поэтому даже спектр разреженного газа являет собой довольно широкие полосы поглощения. Значит, надо и теоретический спектр привести к такой же форме, то есть умножить на аппаратную функцию. Заметим, что свертка теоретического спектра с аппаратной функцией по определению не изменяет ни полного интеграла под результирующей кривой, ни интегралов под отдельными полосами поглощения в спектре. То есть, умножение теоретического спектра на аппаратную функцию не изменяет значений интегральных интенсивностей как отдельных полос поглощения, так и полной интенсивности. Но после такого преобразования сравниваемые полосы обязательно сильно перекроются, а результат их сравнения будет определяться концентрациями компонент в смеси, что и нужно для анализа.

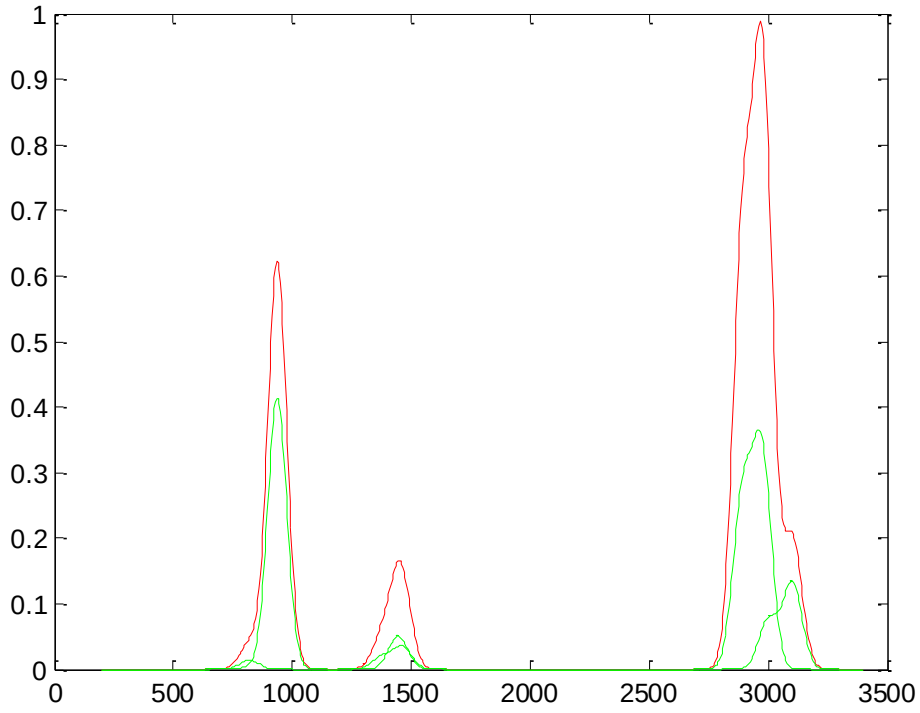
Сказано физиком, сделано программистом. Пересчитаем теоретические спектры этана и этилена, задав всем полосам сравнительно большой параметр полуширины, скажем, 40 см$^{-1}$. Это будет соответствовать регистрации спектра смеси на приборе с весьма посредственным разрешением. После этого проделаем всю вычислительную работу заново.



ИК спектр модели этана при полуширинах полос 40 см$^{-1}$.



ИК спектр модели этилена при полуширинах полос 40 $см^{-1}$.



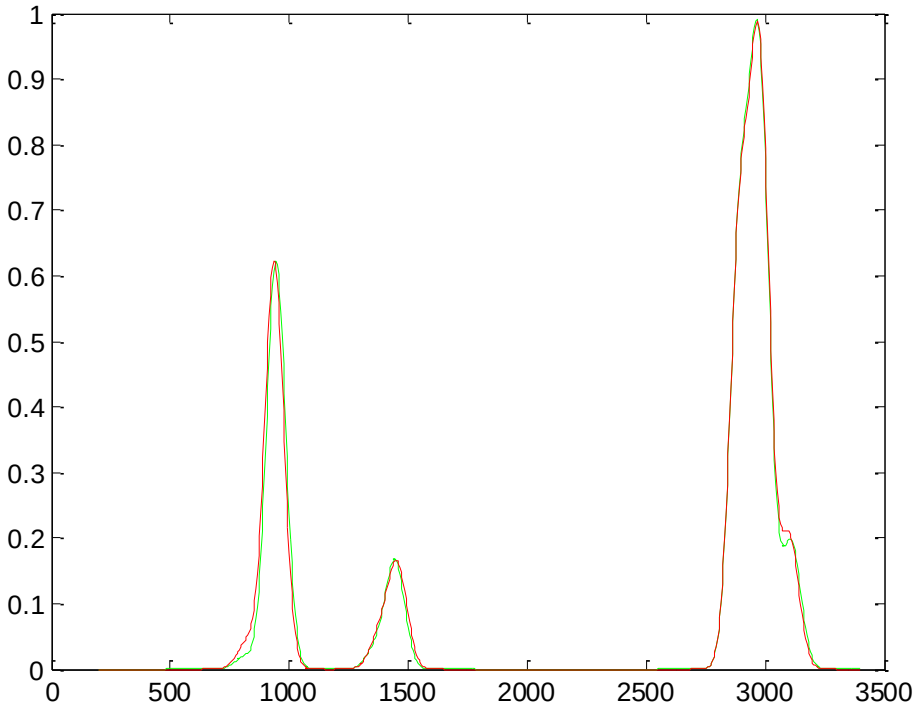
«Экспериментальный» спектр смеси Этан-Этилен показан красным цветом, а входящие в смесь компоненты , в расчете на одну молекулу, – зеленым.

Результаты работы программы анализа

i_best = 123 эту переменную я объясню потом
discrepancy = 0.14
c_bestB = [2.48, 1.39] вместо заданных концентраций [2.5, 1.5]
c1c2 содержит информацию о статистике концентраций в решениях бутстрепа:

	Mean	Std	min	max
$C_{этан}$	2,24	0,32	1,50	2,85
$C_{этилен}$	1,51	0,20	1,14	2,05

Сравним теоретическую кривую, которую предсказывает программа безэталонного анализа, складывая спектры чистых компонент, умноженные на найденные значения концентраци.



«Экспериментальная» кривая дана красным цветом, теоретическая – зеленым.

Просто на глаз видно, что теоретическая кривая, отражающая не только нашу теорию ИК спектров моделей этана и этилена, но и результаты определения концентраций этих веществ в смеси, буквально сливается с «экспериментальной» кривой. Для количественной оценки результатов подсчитывается значение переменной discrepancy. Это сумма квадратов невязок между всеми ординатами двух кривых. Мы видим, что значение discrepancy может служить критерием качества выполненного анализа. В первом случае, когда анализ велся на базе хорошо разрешенных спектров, это значение было велико, а анализ не получился вообще. После внедрения в метод и в программу идеи Грибова о необходимости вести анализ на базе спектров среднего разрешения мы получили почти полное совпадение теории с «экспериментом». Это позволило бы в реальных условиях считать, что анализ дал правильные результаты. Но поскольку мы проводили имитацию анализа смеси с заранее известными нам концентрациями, то мы видим, что точность результатов очень высока. С другой стороны, статистика 200 результатов анализа, полученная путем перебора всех 200 сочетаний вариантов спектров бутстрепа, тоже оказывается полезной. Переменная Std (стандартное отклонение) показывает,

насколько в среднем велик разброс каждой из концентраций. В условиях реального анализа это может помочь оценить возможное влияние неточности молекулярных параметров моделей компонент смеси на точность результатов.

Из проведенного имитационного эксперимента многое проясняется в плане разработки методики анализа. Видно, что для этана погрешность концентрации получается заметно большей, чем для этилена. Мы можем в этом разобраться, если заметим, что все полосы поглощения этана перекрываются с полосами этилена, тогда как у этилена имеются изолированные от этана полосы. Дальнейший осмотр особенностей эксперимента пусть проведут химики-аналитики, для которых мы и стараемся.

Наконец, о переменной i_best. В первом, неудачном примере анализа, она была равна 65. Во втором, удачном примере анализа, она была равна 123. Это номер испытания бутстрепа, одинаковый и для этана, и для этилена, который дан наилучший по критерию discrepancy результат анализа из 200 таких проб.

Программа nonetalon_n.m в поисках наилучшего решения пробегает все варианты бутстрепа, придавая один и тот же номер случайным спектрам и этана, и этилена. Из теории анализа ясно, что наилучшее решение должно получиться тогда, когда под одним и тем же номером окажутся два спектра компонент, наиболее близкие к неизвестным нам спектрам чистых соединений. Бутстреп для того и варьирует молекулярные параметры наших моделей, что он надеется на удачу – совершенно случайно оба спектра окажутся именно такими, очень близкими к реальным. Но вероятность такой удачи очень мала. Потому мы и заставляем бутстреп генерировать нам очень много случайных спектров, 200 вариантов для каждого компонента.

Вот мы и видим, что в первом примере наименьшая невязка спектров получилась для сочетания случайных спектров этана и этилена с номерами 65 и 65. При сравнении уширенных спектров наименьшая невязка получилась для сочетания случайных спектров этана и этилена с номерами 123 и 123. Естественно и результаты анализа по этим двум случайным сочетаниям спектров получились кардинально различными. Ясно, что пара спектров с номерами 123 и 123 дает лучший результат анализа.

Начинаем соображать, что мы могли бы выловить и еще более удачный результат анализа, если бы нашли более удачное сочетание в другой паре спектров, с разными номерами в двух независимых коллекциях бутстрепа. Но тогда надо иначе организовывать поиск. Надо бы перебрать все сочетания спектров из коллекции спектров этана и из коллекции этилена. Таких сочетаний 200*200 = 40000. С таким перебором, с направленным поиском лучшего сочетания спектров персональный компьютер еще может справиться. Но мы нацелились на анализ многокомпонентных смесей. И тут с полным перебором не справится и самый мощный суперкомпьютер.

Мы придумали и реализовали алгоритм полунаправленного поиска, который значительно быстрее полного направленного поиска. Но в данной книге мы не будем описывать этот алгоритм и его реализацию в программе ne_2plus.m. Текст этой программы доступен для ознакомления в документации, прилагаемой к данной книге. Также не будем приводить результаты анализа многокомпонентных смесей, проведенного с помощью этой

программы. Если кто-то из читателей заинтересуется безэталонным анализом, то мы можем его познакомить с нашими специальными публикациями.

Предварительные выводы на основе глав 2-4

Очень серьезных выводов пока сделать нельзя. Не позволяет принятый стиль компоновки материала и написания книги. Пока я только обращаюсь к потенциальным читателям с предложением – я вам покажу, как я это делаю. А вы решите, будете ли вы делать также, и будете ли вообще.

С другой стороны, в предисловии я обещал показать пользу от визуального моделирования колебаний молекул. Мне кажется, что в главах 3 и 4 я кое-что уже показал.

Долгая история развития теории колебаний молекул в нашей стране была очень неровной. К 70-м годам прошлого века она подошла весьма энергично, предложив химикам две замечательные возможности.

1. Теоретическая интерпретация колебательных спектров сложных органических соединений становилась существенным блоком при построении экспертных систем, предназначенных для распознавания неизвестных соединений. Экспертные системы опирались на данные самых разных видов спектроскопии, но борьбу с многозначностью ответа брала на себя теория ИК спектров. Те варианты структур неизвестного соединения, которые не давали правильный прогноз ИК спектра, решительно отбрасывались.
2. Возможность так уточнить параметры модели путем решения обратных спектральных задач, чтобы расчетный спектр был очень близок к экспериментальному, позволяла химикам пристально всматриваться в значения полученных параметров и делать углубленные выводы о внутренней электронной жизни исследуемого органического соединения. То есть интерпретировать и прогнозировать реакционные способности молекул.

Затем наступили другие времена. Оказалось, что радиочастотные резонансные спектральные методы заметно обгоняют по своим возможностям ИК спектроскопию в плане распознавания структур органических соединений. Экспертные системы стали базироваться на ЯМР спектрах.

Мощные квантово-химические программы позволяют значительно проще рассмотреть особенности электронного строения сложных органических молекул.

Было впору прекратить дальнейшую разработку разделов теории колебаний молекул.

Однако жизнь не стоит на месте. Нашлись новые области применения теории и методов расчета колебаний молекул. Более того, новые приложения потребовали новых усилий от теоретиков и разработчиков специализированных программ. Что нам, теоретикам и разработчикам программ приятно.

В главе 3 я показал, как прогноз картины простых нормальных колебаний и сложных сумм этих колебаний помогает обдумывать возможность или невозможность

прохождения химической реакции. В последующих главах я покажу, как надо действовать, чтобы перейти к строгому квантовому решению вопроса о вероятности предполагаемой реакции. И станет понятно, что еще очень далеко до того момента, когда появятся сервисные программы, дающие мгновенно ответ на такой вопрос. Пока нащупана сфера деятельности, а сделать теоретикам и программистам предстоит еще очень много.

Здесь, в главе 4 я показал новую область применения методов априорного расчета спектральной кривой ИК поглощения органических молекул. Это безэталонный спектральный анализ. Возврат к теории ИК спектров не случаен. Физика молекул пока научилась выполнять инженерный расчет интенсивностей только в ИК спектрах, опираясь на надежные колебательные модели ранее исследованных соединений. Другие виды спектроскопии пока здесь проигрывают.

В следующих главах я покажу и другие новые области применения техники расчета колебаний сложных молекул. В частности, в механохимии. Так что, я постепенно выполняю обещания, данные в предисловии.

Хуже обстоит с выводами по главе 2. Подозреваю, что потенциальный читатель, пройдя вместе со мной весь путь от геометрии аллена до прогноза его спектральной кривой ИК поглощения, скажет – как это трудно и нудно. Особенно воспроизводить чужие модели и расчеты. И будет прав.

Теперь моя задача показать потенциальному читателю, что имеется множество способов упростить и облегчить процесс визуального моделирования колебаний молекулы. Конечно, в принятом режиме работы с книгой (я показываю, как я делаю, а вы решаете, будете ли так делать вы, и будете ли вообще) освоить эти способы невозможно. Но можно предварительно осмотреть и решить, что стоит ознакомиться и с подробной инструкцией ко всем основным программам. Такая инструкция приведена в конце книги. Соединив инструкцию с методическими советами, разбросанными в тексте книги, можно выработать свои собственные способы упрощения работы при визуальном моделировании колебаний молекул.

Я же выдвигаю тезис:

Чем сложнее и обширней структура органической молекулы, тем проще построить ее колебательную модель, пользуясь средствами системы LevInfinite/

В следующей главе я постараюсь этот тезис доказать.

Литература

1. В.А. Дементьев, А.В. Сорока, Т.Г. Химочко. Особенности применения метода бутстрепа при нахождении сложных статистических функций от малых выборок в биологических и медицинских исследованиях. Биомедицинская химия, Том 50, Приложение № 1, ГУ НИИ биомедицинской химии РАМН, М., 2004, с. 117-126.
2. О.Е. Родионова. Тезисы докторской диссертации, М., 2007.

3. Л.А. Грибов, В.А. Дементьев, А.Т. Тодоровский. Интерпретированные колебательные спектры алканов, алкенов и производных бензола. Наука, М., 1986, 495 с
4. Л.А. Грибов, М.Е. Эляшберг, В.И. Баранов. Безэталонный спектральный анализ. Изд. УРСС, М., 2002, 317 с.